



Analyser le mediascape : retour sur un projet pluridisciplinaire

Julien Velcin
Laboratoire ERIC
Université de Lyon
julien.velcin@univ-lyon2.fr
[@jvelcin](#)

Outline of my talk

- 1) Exploring the informational landscape
- 2) Topic modeling as a major tool in NLP
- 3) Studying the mediascape
- 4) A « user in the loop » perspective
- 5) Conclusion in a DH point of view

Outline of my talk

- 1) Exploring the informational landscape**
- 2) Topic modeling as a major tool in NLP
- 3) Studying the mediascape
- 4) A « user in the loop » perspective
- 5) Conclusion in a DH point of view

home

headlines

Thursday
28 January 2016

Now 14°C

17:00 12°C 	20:00 9°C 	23:00 8°C 	02:00 8°C
-------------------	------------------	------------------	------------------

Lyon

Zika virus spreading 'explosively', says World Health Organisation

Director general convenes emergency committee saying it is deeply concerning virus linked to birth defects has now been detected in more than 20 countries



244

Brazil Recife, city at centre of Zika epidemic

Video What you need to know

'Should I cancel my holiday?' Latest advice for travellers



Denmark PM's tough stance criticised by international media but has popular support at home

Immigration Sweden sends sharp signal with plan to expel up to 80,000 asylum seekers

4 December 2015

THE HUFFINGTON POST

UNITED KINGDOM

Edition: UK

Search The Huffington Post

Like

632k

Follow

Follow

424k

FRONT PAGE NEWS POLITICS BUSINESS TECH YOUNG VOICES COMEDY ENTERTAINMENT CELEBRITY LIFESTYLE PARENTS BLOGS

Politics • COP21 • Building Modern Men • What's Working • Environment • Media • Women • Impact • Entrepreneurs • Young Talent • Christmas • Smart Living



#malavita

Top

Direct

Comptes

Photos

Vidéos

Autres options ▾

Suggestions · Actualiser · Tout afficher

**Kaplan International** @k... ×[Suivre](#) Sponsorisé**Aras BOZKURT** @arasbozkurt ×[Suivre](#)**Stéphane Pouyllau** @spouyl... ×[Suivre](#)

Trouver des amis

Tendances · Modifier

#EnVoiture

Sponsorisé par Allianz France

#ASSEPSG

#JacquelineSauvage

#Camping

#TheVoice

#Malavita

Milan

Bordeaux

Ronaldo

Florian Thauvin

Benoît Violier

11 nouveaux résultats

**Marion** @Marion_LeJct · 2 min

Trop cool ce film de #LucBesson #Malavita #TF1

**Gabi** 🙌 @11gabi_01 · 2 min

Et putain... 🙌 #Malavita

**Mouna Camara** @mouna_camara · 2 min

Tres bon film #Malavita

**Stephanie L** @SLidouren · 2 min

Après #Malavita piace à #LOLUSA

**Black Mamba** @SmallHawkeye · 2 minBon ce film était bof. Rien d'exceptionnel, des petits passages marrant.
Dommage avec un tel casting...

#Malavita

1

**Ree** @HirtRee · 2 min

Film d'action américain en normandie 🤩 #Malavita

Exploring textual corpora

- Top-down approaches
 - regexp (e.g., “terror[a-z]*”)
 - keyword-based queries
 - (supervised) classification schemas
- Bottom-up approaches
 - basic statistics
 - projection to low-level spaces (e.g., PCA)
 - structuration by (co-)clustering or topic modeling
 - More advanced summarization techniques (e.g., metromaps)
- Mixed approaches
 - weakly-supervised clustering
 - active learning schemas

Outline of my talk

- 1) Exploring the informational landscape
- 2) Topic modeling as a major tool in NLP**
- 3) Studying the mediascape
- 4) A « user in the loop » perspective
- 5) Conclusion in a DH point of view

Some background

- Bag-of-words assumption
- Usual preprocessing
 - remove numbers and punctuation
 - remove stopwords
- Classic input:



Terms	Docs																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
data	1	1	0	0	2	0	0	0	0	0	1	2	1	1	1	0	1	0	0	0
examples	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
introduction	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
mining	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0
network	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1
package	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0

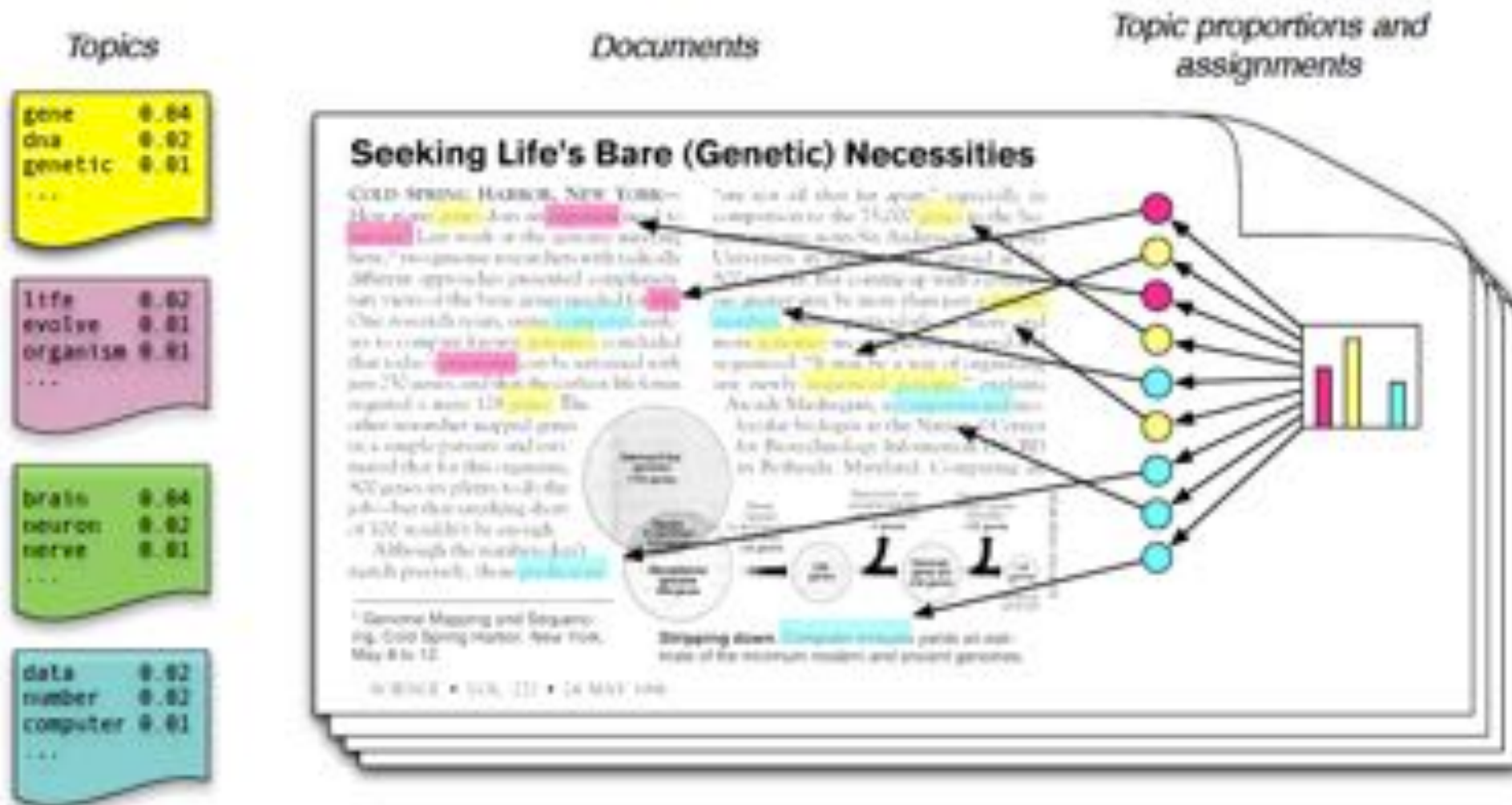
Why topic learning?

Topic modeling provides methods for automatically organizing, understanding, searching, and summarizing large electronic archives:

- discover the **hidden themes** that pervade the collection
- **annotate** the documents according to those themes
- use annotations to organize, summarize, and **search** the texts

(some slides are taken from the talk of D. Blei for KDD 2011)

Discovering latent structures



(Blei et al., 2003)

ReadiTopics (Velcin et al., 2018)

Learn less

ReadiTopics is a visualization application developed at the University of Lyon with a close collaboration with Montpellier and TETIS. It allows a user to browse the results of a topic model and find the best possible label to get an understanding of the topics' content.

Pick a dataset:

HP

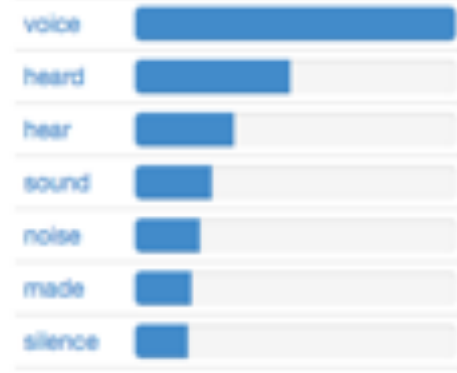
Pick a topic:

Topic 45

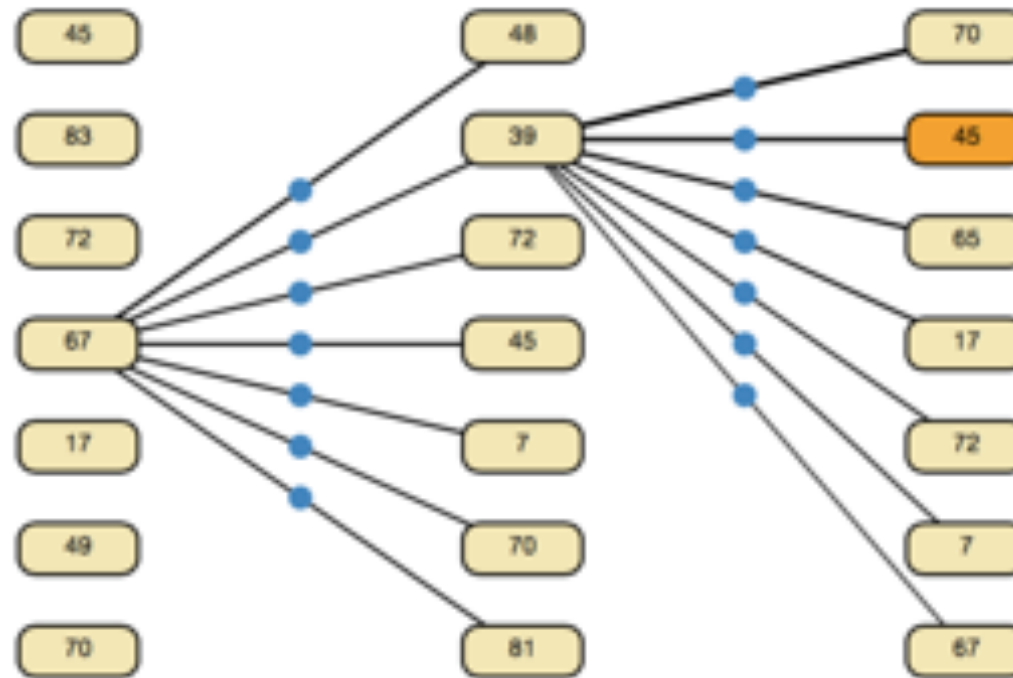
Word per topic:



Top words for topic 45 of dataset HP



Current **topic 45** : voice, heard, hear, sound, noise... (847.8) [hide topic](#)



Hidden topics:

10 18 19 22 23 28 35 59 61 95 96

Top-10 documents [show](#)

doc *Harry_Potter_7.txt-3523* 95 39 62 28 22

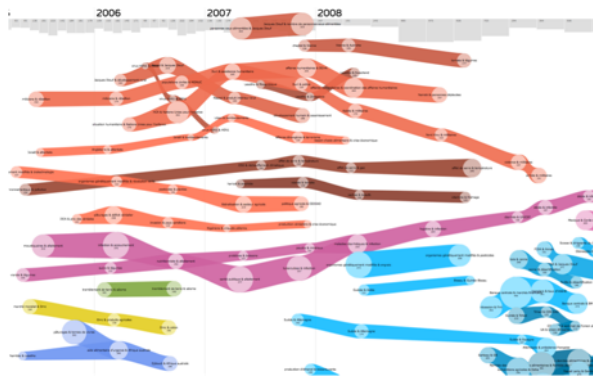
'We've already got an Invisibility Cloak, ' said Harry. 'And it's helped us rather a lot, in case you hadn't noticed!' said Hermione. 'Whereas the wand would be bound to attract trouble-- ' 'Only if you shouted about it, ' argued Ron. 'Only if you were prat enough to go dancing around waving it over your head, and singing. I've got an unbeatable wand, come and have a

Outline of my talk

- 1) Exploring the informational landscape
- 2) Topic modeling as a major tool in NLP
- 3) Studying the mediascape**
- 4) A « user in the loop » perspective
- 5) Conclusion in a DH point of view

What's a « mediascape »?

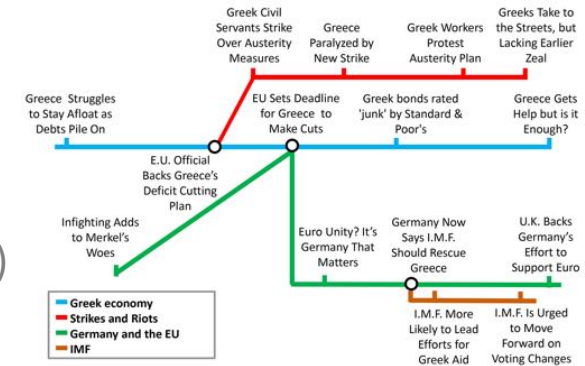
Appadurai A.: Disjuncture and Difference in the Global Cultural Economy, Theory Culture Society 1990; 7; 295



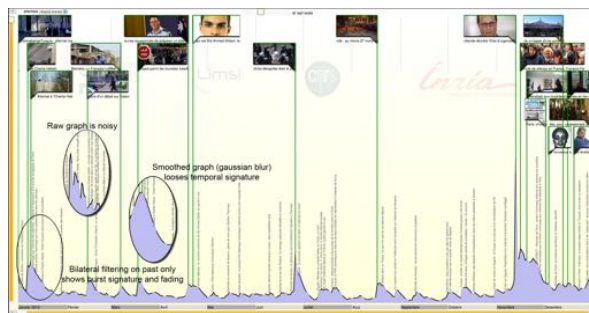
<http://pulseweb.cortex.net>

Projet Pulseweb
(Cointet, Chavalarias...)

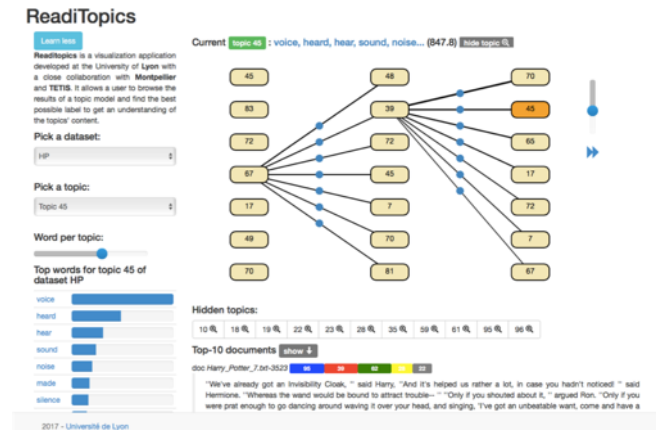
Metromaps
(Shahaf et al., 2015)



Readitopics
(Velcin et al., 2018)



Chronolines
(Nguyen et al., 2014)



<https://github.com/Erwangf/readitopics>

What we did

- Joint work with a sociologist (J.C. Soulages, Max Weber lab), in a project related to data journalism (Velcin et al., workshop @EGC 2017)
- As input: a collection of documents (here, newspapers from the **Huffington Post**)
- As output: distribution over topic categories
- Two levels of categories:
 - basic level, found by the topic model (here, LDA)
 - high level, labeled by experts (here, J.C. Soulages and partners from Brazil)

Comparing news media

- Usual preprocessing (tokenisation, stopwords...)
- Three versions of the same media (HuffPost):

Version	langue	#articles	longueur	#mots
US	anglais	12 067	454.4	5 482 661
FR	français	4 133	369.6	1 527 416
BR	portugais	2 355	429.5	1 011 373

- How to compare those three versions by using LDA?
 - > associate each topic with one **given** category (e.g., sport or media)
 - > up to now, this is manual!
- Estimate the **importance** of every category (here, volume of words tagged by the covered topics)

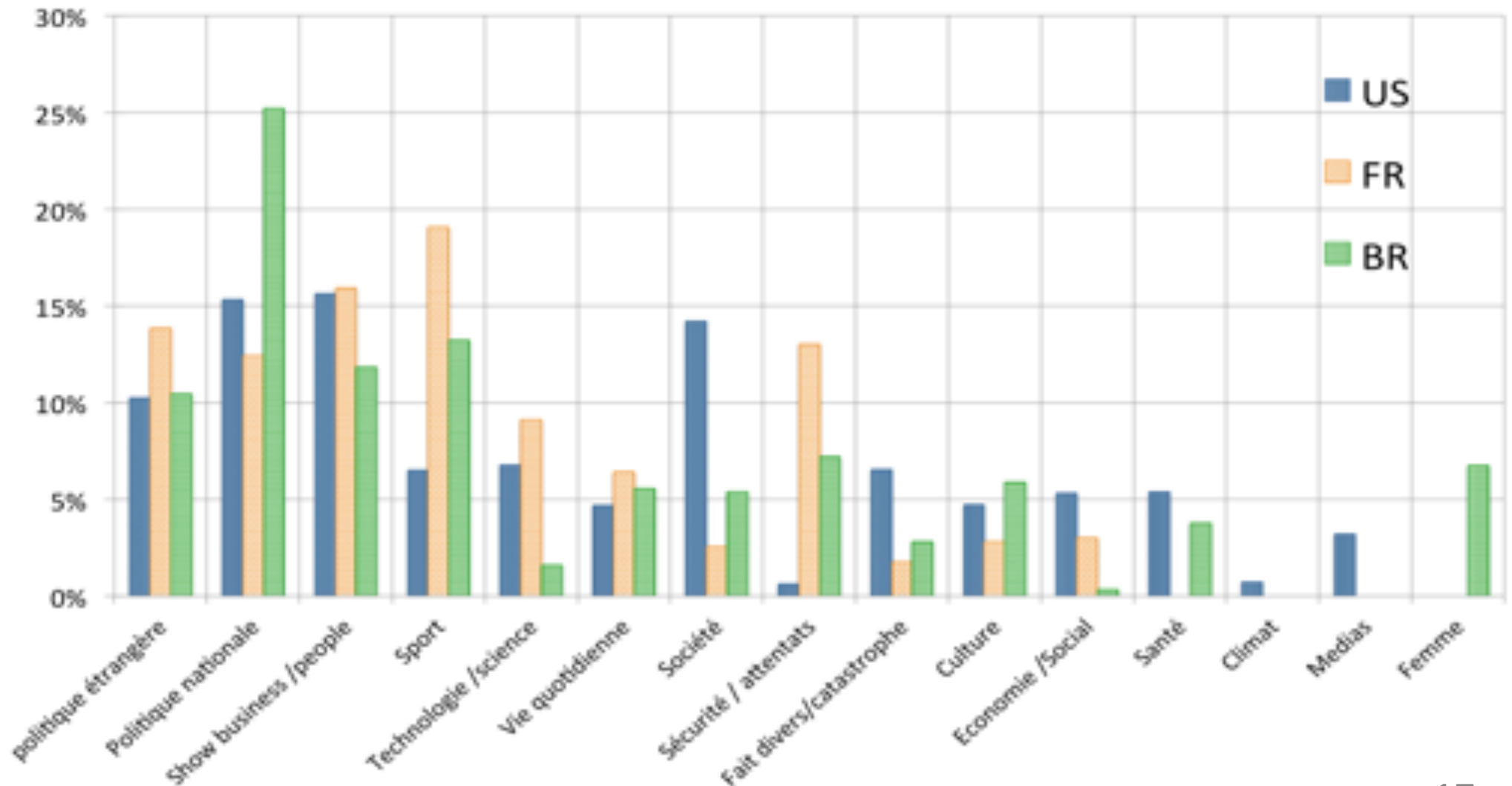
Some topics extracted by LDA

en français (sur 4133 articles) :			
topic	#doc	cat.	mots les plus probables
z ₁₈	28	1	manifestation, paris, police, travail, loi, contre, syndicats, place, bastille, 2016
z ₁₉	36	1	loi, travail, gouvernement, l'état, texte, l'assemblée, d'urgence, mois, projet, conseil
z ₂₅	39	2	jeux, rio, olympiques, olympique, août, jo, athlètes, 2016, brésil, cérémonie
z ₄₇	18	3	morandini, jean-marc, inrocks, catherine, l'animateur, lui, qu'il, europe, comédiens, plainte
z ₇₃	47	4	nice, 14, l'attentat, anglais, promenade, camion, attentat, police, soir, christian
en anglais (sur 12067 articles) :			
z ₁₄	92	5	refugees, children, refugee, people, countries, world, syrian, rights, million, year
z ₂₁	74	2	gymnastics, biles, olympic, team, simone, olympics, gymnast, gold, rio, hernandez
z ₃	46	6	pokemon, game, pokémon, playing, players, catch, «pokemon, go», pizza, play
z ₅₀	56	7	muslim, religious, muslims, faith, church, god, christian, religion, hate, american
z ₂₇	140	8	clinton, voters, trump, poll, polls, americans, election, support, vote, relationships
en portugais (sur 2355 articles) :			
z ₄₄	52	8	dilma, presidente, impeachment, senado, senadores, processo, senador, rousseff, julgamento, defesa
z ₅₈	7	9	sexo, menstruação, durante, rao, mccane, comédia, realmente, corpo, riso, menstruada
z ₇₁	11	7	negros, brancos, negras, pessoas, racial, negra, racismo, país, movimento, black
z ₃₇	57	2	brasil, vôlei, jogo, medalha, vitória, ouro, seleção, set, brasileiras, torcida
z ₉₉	20	7	lgbt, gay, preconceito, violência, sexual, direitos, família, orgulho, estupro, aborto

Les catégories attribuées ici (cat.) correspondent à : 1- Economie / Social, 2- Sport / JO, 3- Show business / people, 4- Sécurité / attentats, 5- Politique étrangère, 6- Technologie / science, 7- Société, 8- Politique nationale, 9- Santé.

Compared results

Normalized distribution over the 15 categories
(remember that each category can be associated to **multiple** topics)



Outline of my talk

- 1) Exploring the informational landscape
- 2) Topic modeling as a major tool in NLP
- 3) Studying the mediascape
- 4) **A « user in the loop » perspective**
- 5) Conclusion in a DH point of view

Human in the loop for topic models

- Give more insight about the produced structure
 - work on topic coherence (Röder et al., 2015)
 - topic labeling (Mei et al., 2007) (Gourru et al., 2018)
- Let the user interfere with the model
 - interactive topic modeling (Hu et al., 2014)
 - make the model selection easier
- **Towards learning the (meta) categories**
 - semi-supervised, active learning techniques
 - **refine the categorical structure with humans!**

Newsbrowsers

The screenshot shows a news browser application with a sidebar on the left and a main content area. The sidebar contains a table with columns 'Fichier' and 'Importés', showing a file named 'fr-fev-juin_cli...' with 563/563 items imported. The main content area has a top navigation bar with tabs: 'Recherche', 'Classifier périodes', 'Requête', 'Classification', and 'Topics'. Below this are several search filters: 'Mot clé', 'Tag manu...' (with a 'Sélectionner' button), 'Tag auto', 'Période', and 'Statut' (set to 'Non traité'). A 'Recherche' button is located below these filters. The main content area displays a table of news items with columns 'Id', 'Titre', and 'Auteur'. A right-hand panel is open, showing a 'Créer tag' field, a 'Chercher' field, and a list of tags including 'AGRICULTURE', 'ALIMENTATION', 'AUTOMOBILE', 'BIODIVERSITE', 'CLIMAT', 'CLIMATOSCEPTIQUE', 'COLLECTIF', 'DECHETS/RECYCLAGE', 'DENEIGATION', 'DENONCIATION', 'DEVELOPPEMENT DURABLE', 'DOUBLOON', and 'ECOLOGIE'. At the bottom of the application, there is a status bar with the following text: '[09:36:49] Ligne 1 : Ligne commençant par titre (en-tête)', '[09:36:49] Count success read = 563', and '[09:36:49] Nombre de tags importés : 28'. On the far right, there is a system tray area with a 'Selection' button and a 'Updates ready to...' notification.

Id	Titre	Auteur
	L'anecdote adorable de Catherine Laborde sur s...	(no author)
	Face au projet de Tour Triangle, les écologist...	(no author)
	Devenu un symbole des anti-Donald Trump, Ollie...	(no author)
	Les bouteilles de lait en plastique seront-ell...	(no author)
	François Hollande n'a donné aucun conseil de c...	(no author)
	Petit tour du monde des sources d'énergie reno...	(no author)
	Comment débrancher François Fillon? Les quatre...	(no author)
	Voici ce qu'il faut savoir pour (bien) recycle...	(no author)
	Il n'est pas trop tard pour faire barrage au CETA	(no author)
	La réaction de Donald Trump à l'attaque au mus...	(no author)
	Le gouvernement suédois signe une loi sur l'en...	(no author)
	Chronique d'une victoire annoncée - Episode 3:...	(no author)

Outline of my talk

- 1) Exploring the informational landscape
- 2) Topic modeling as a major tool in NLP
- 3) Studying the mediascape
- 4) A « user in the loop » perspective
- 5) **Conclusion in a DH point of view**

Take-home message

- For studying the informational landscape:
 - semi-supervised / active learning is the key
 - clear idea of the targeted application
 - lots of engineering work (needs funding)
- Collaboration with LLSSH needs:
 - mutual understanding
 - trust
 - time
- And by the way...
 - life-cycle of information (Davidson et al., 2020)

References

(Appadurai A., 1990) Appadurai A.: Disjuncture and Difference in the Global Cultural Economy, **Theory Culture Society**, 1990; 7; 295.

(Blei et al., 2003) Blei, D. M., Ng, A. Y., & Jordan, M. I.: Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022, 2003.

(Davidson et al., 2020) Davidson I., A. Gourru, J. Velcin and Y. Wu: Behavioral differences: insights, explanations and comparisons of French and US Twitter usage during elections. *Social Network Analysis and Mining (SNAM)*, 10: 6.

(Gourru et al., 2018) Gourru, A., Velcin, J., Roche, M., Gravier, C., & Poncelet, P.: United we stand: Using multiple strategies for topic labeling. In *International Conference on Applications of Natural Language to Information Systems (NLDB)*, pp. 352-363, 2018, Springer.

(Velcin et al. 2017) Velcin J., J.C. Soulages, S. Kurpiel, D.L. Otavio, M. Del Vecchio et F. Aubrun: Fouille de textes pour une analyse comparée de l'information diffusée par les médias en ligne : une étude sur trois éditions du Huffington Post. **Atelier Journalisme computationnel @EGC**, 2017.

(Velcin et al., 2018) Velcin, J., Gourru, A., Giry-Fouquet, E., Gravier, C., Roche, M., & Poncelet, P.: Readitopics: make your topic models readable via labeling and browsing. *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI 2018, demo track)*, pp.5874-5876, Stockholm, Sweden.